The Leadership Quarterly 30 (2019) 101340

Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

When are women willing to lead? The effect of team gender composition and gendered tasks

Jingnan Chen[a,*], Daniel Houser[b]

[a] Economics Department, Business School, University of Exeter, Exeter EX4 4PU, United Kingdom
[b] Interdisciplinary Center for Economic Science (ICES) and Department of Economics, George Mason University, Fairfax, VA 22030, United States of America

ARTICLE INFO

Keywords:
Gender diversity
Team performance
Stereotype
Board

ABSTRACT

It is a well-documented phenomenon that a group's gender composition can impact group performance. Understanding *why* and *how* this phenomenon happens is a prominent puzzle in the literature. To shed light on this puzzle, we propose and experimentally test one novel theory: through the salience of gender stereotype, a group's gender composition affects a person's willingness to lead a group, thereby impacting the group's overall performance. By randomly assigning people to groups with varying gender compositions, we find that women in mixed-gender groups are twice as likely as women in single-gender groups to suffer from the gender stereotype effect, by shying away from leadership in areas that are gender-incongruent. Further, we provide evidence that the gender stereotype effect persists even for women in single-gender groups. Importantly, however, we find that public feedback about a capable woman's performance significantly increases her willingness to lead. This result holds even in male-stereotyped environments.

Introduction

The gender composition of teams, and how it impacts organizational outcomes, has attracted increasing attention in the media and the leadership literature. Recently, for example, people have heatedly debated the benefits of increasing the female presence on boards, and the merits of gender diversity in leadership[1]. It is well-substantiated that female and male leaders differ systematically in their core values, leadership style and risk attitudes (cf., Adams, Funk, Barber, Ho, & Odean, 2012; Druskat, 1994; Eagly, Makhijani, & Klonsky, 1992). The extant literature has yet to reach a consensus on the causal effects of the gender diversity of corporate boards on firms' performance, with some studies yielding positive results, and others producing null or negative outcomes (e.g., Eagly, 2016; Yang, Riepe, Moser, Pull, & Terjesen, 2019). It is worth noting that some benefits of greater female leadership include female leaders as role models for fellow aspiring women in the organizations (cf., Arvate, Galilea, & Todescat, 2018; Gilardi, 2015).

The call for gender diversity is especially loud within male-dominated and traditionally male-stereotyped industries, such as the technology industry. A special report by the U.S. Equal Employment Opportunity Commission (EEOC) highlighted the technology sector as having particularly "concerning trends" in employment, despite being a major source of economic growth. The technology sector employed a significantly smaller share of women compared to overall private industry (36% in technology and 48% overall in private industry). The fact that women are underrepresented in a male-dominated high earning industry is hardly surprising. Similar patterns were observed in political sphere where women are underrepresented in legislative bodies (cf., Kanthak & Woon, 2015). Gender stereotypes, especially stereotype-based expectations of inferiority, are considered to be the major factors contributing to the absence of gender diversity and underrepresentation of women, especially in leadership roles. Gender-based expectations, founded in stereotype bias, can impact not only who people regard as "fitting" for leadership roles, but also a person's willingness to lead (Eagly & Karau, 2002; Hoyt & Blascovich, 2010; Hoyt & Murphy, 2016).

The effect of gender composition on organizational performance and group decision-making has been well documented using both observational and experimental data. Intriguingly, this effect persists even *after* controlling for observable characteristics of individuals (Apesteguia, Azmat, & Iriberri, 2012; Azmat & Petrongolo, 2014; Bagues, Sylos-Labini, & Zinovyeva, 2017; Berge, Juniwaty, & Sekei, 2016; Hoogendoorn, Oosterbeek, & Praag, 2013; Joecks, Pull, & Vetter, 2013; Kirsch, 2018; Terjesen, Sealy, & Singh, 2009). Nevertheless, the

* Corresponding author.
E-mail addresses: j.chen2@exeter.ac.uk (J. Chen), dhouser@gmu.edu (D. Houser).

[1] See, for example, an opinion piece in the Huffington Post, http://www.huffingtonpost.com/caroline-turner/gender-diversity-on-boards_b_7744588.html; see also the Philadelphia Business Journal, http://www.bizjournals.com/philadelphia/news/2017/03/28/mentoring-mondayhow-gender-diversity-in-the.html.

https://doi.org/10.1016/j.leaqua.2019.101340
Received 5 June 2018; Received in revised form 8 October 2019; Accepted 9 October 2019
Available online 20 November 2019

literature has yet to solve the puzzle of *how and through which channels* gender composition affects group performance. This question is crucially important to academics, organizations, and policy makers. To formulate appropriate policy interventions, we must shed light on the underlying mechanisms at work.

In this paper, we take a step towards filling the gap in the literature. We design and implement laboratory experiments to test an important potential mechanism: as a result of gender stereotyping, the gender composition of a group may moderate one's willingness to lead a group. For example, if, in gender-diverse groups, relatively lower-skilled men are more likely to lead than higher-skilled women, this could detrimentally impact the quality of the group's output (holding other group members' ability constant).

Gender stereotypes are widely held in society. According to stereotype threat theory (Steele & Aronson, 1995), stereotype boost theory (Shih, Pittinsky, & Ho, 2011), and role congruity theory (Eagly & Karau, 2002), people are expected to perform better when characteristics required for a task are congruent with gender stereotypes (positive stereotypes) about their social group (e.g., men are more proficient at mathematical tasks). By contrast, they are expected to perform worse when these characteristics are incongruent with gender stereotypes (negative stereotypes) about their social group (e.g., women are less proficient at mathematical tasks). We denote this effect the *gender stereotype effect* (GSE). Coffman (2014) demonstrated empirically across decision domains with varying gender stereotype that *when holding ability constant*, women (men) are less likely to put forward their answers as the group lead answers to male (female) stereotyped problems. Inefficiency is the potential negative consequence, stemming especially from under-contribution by equally able women in male-stereotyped domains, and equally competent men in female-stereotyped domains.

In this paper, we explore *why* a group's gender composition is likely to influence its performance through GSE. By varying the comparison set, a group's gender composition could impact the salience of one's gender identity and her corresponding gender stereotype (Cohen & Swim, 1995; Cota & Dion, 1986; Hoyt, Johnson, Murphy, & Skinnell, 2010). For example, when a woman is in an all-female group, her female gender identity may become less salient and she may suffer less from GSE, in that her willingness to lead the group may be less influenced by the gender stereotype of the decisions. As a result, she may be equally likely to take the lead and offer her qualified ideas to both male and female stereotyped tasks and improve the overall group performance. By contrast, if a woman is placed in a majority-male group, her female identity may become more salient and she may hold back when confronted with male-stereotyped problems. Not only is the overall quality of the group's ideas reduced, but a woman (man) in a male (female) dominated group may be more likely to be overlooked for promotion or advancement opportunities as a consequence of shying away from providing her talents to the team.

Our contributions are fourfold: First, we offer and empirically test a novel mechanism for the gender composition effect on willingness to lead and team performance using tasks from different decision domains. Ours is one of the first studies to bring together insights from psychology, management, leadership studies and economics, and provide new evidence on why gender composition affects performance. Second, we make a significant methodological contribution to the literature. Much of the previous observational research about the gender composition effect on performance in management and applied psychology has been correlational and not causally identified (cf., Antonakis, Bendahan, Jacquart, & Lalive, 2010; Haslam, Ryan, Kulich, Trojanowski, & Atkins, 2010; Miller & Del Carmen Triana, 2009; Smith, Smith, & Verner, 2006). Even when laboratory experiments have been used to establish the causality, measures of the dependent variables have not been consequential or incentivized (cf., Heilman & Haynes, 2005). In particular, one of the main techniques used in the literature to activate GSE—gender priming—is currently under debate due to replicability and experimenter demand concerns (Cesario, 2014; Doyen, Klein, Pichon, & Cleeremans, 2012; Flore, Mulder, & Wicherts, 2019; Lonati, Quiroga, Zehnder, & Antonakis, 2018). In contrast to previous research, we exogenously and subtly vary the gender composition of the group to trigger GSE. Likewise, all decisions made in our experiment are adequately incentivized. In so doing, we minimize the experimenter demand effect and offer an ecologically valid measure of the outcome variables, while still enabling rigorous inference regarding the causal link between gender composition, willingness to lead, and performance. Third, we offer important new evidence that public performance feedback effectively encourages qualified women to lead, even in male-typed environments. Finally, our study contributes to the literature on leader emergence. Whereas previous key contributions have focused on the personality traits of leaders (e.g., Judge, Bono, Ilies, & Gerhardt, 2002), we focus instead on features of the environment (particularly gender stereotype and gender composition) that shape incentives for individuals to pursue leadership (Zehnder, Herz, & Bonardi, 2017).

The remainder of the paper is organized as follows: Literature review reviews the relevant literature; Experimental design and procedures describes the experimental design and procedure; Predictions outlines our predictions; Results presents the main results; Discussion discusses and Conclusion concludes.

Literature review

Instrumental leadership

Our study focuses on functional and instrumental leadership behavior (sometimes equated to pragmatic leadership) (Antonakis & House, 2014; Lord, 1977; Morgeson, DeRue, & Karam, 2010; Mumford & Van Doorn, 2001), as compared to transactional or charismatic and transformational leadership (Bass, 1985; Burns, 1978). For a comprehensive review of leadership styles, see Anderson and Sun (2017).

Following Antonakis and House (2014 p.749), instrumental leadership is “the application of leader expert knowledge on monitoring of the environment and of performance, and the implementation of strategy and tactical solution.” Fleishman et al. (1991) point to the responsibility of the leader in problem solving, and suggest that to be effective, a leader must be equipped with problem-solving skills and expert knowledge (Connelly et al., 2000; Morgeson et al., 2010). The problem-solving role of leaders seems especially critical when a team faces complex tasks (French & Raven, 1968).

In economics, there is growing interest in studying leadership, although leadership has yet to become an established subfield. As elaborated by Zehnder et al. (2017), the leadership literature in psychology and management has been largely running in parallel to the leadership literature in economics. We are one of the first studies to bridge those fields. We use laboratory economics experiments to capture instrumental leadership, and more importantly, the willingness to lead. To the best of our knowledge, we are one of the first studies to investigate one's willingness to lead using an incentive-compatible elicitation method. A topic in the literature closely linked with willingness to lead is leadership aspiration, which is shown to predict leader emergence (see, e.g., Reitan & Stenberg, 2019). In our environment, group members coordinate to solve problems from different fields. Consequently, the most suitable leader should be the one with the greatest expertise in the subject area. Using incentivized elicitation tasks, we quantify the degree to which capable members were willing to step forward to lead. Crucially, our experiment design enables us to demonstrate how their willingness to lead is influenced by a group's gender composition.

Gender stereotype effect

Over the past two decades, substantial research in psychology has investigated the effect of stereotype on performance. There are two strands of theory on this topic: stereotype threat theory and stereotype

boost theory. Stereotype threat theory predicts that negative stereotypes hurt performance, while stereotype boost theory predicts that positive stereotypes will boost performance. Empirical studies of stereotype threat generally find that negative stereotypes undermine performance of stereotyped individuals (e.g., academic performance, as well as performance in other domains, including athletic and memory tasks). Women, individuals with lower socioeconomic status, and the elderly are often highly detrimentally impacted by stereotype threat (cf. Aronson, Quinn, & Spencer, 1998; Chasteen, Kang, & Remedios, 2011; Croizet & Claire, 1998; Levy, 1996; Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995; Stone, Sjomeling, Lynch, & Darley, 1999). A large body of work also shows that activating positive stereotypes can help boost performance (cf. Kray, Thompson, & Galinsky, 2001 for a review, see Shih et al., 2011 Spencer et al., 1999). Mechanisms thought to account for stereotype threat and stereotype boost include changes in stress and anxiety, the mediation in self-efficacy, beliefs about one's own ability (self-doubt/self-confidence) and changes in neural processing efficiency (Shih et al., 2011).

Our study focuses on gender stereotypes. Following the stereotype literature and role congruity theory (Eagly & Karau, 2002), people experience gender stereotype threat when performing tasks with characteristics that are incongruent with gender stereotypes about their social group, namely, tasks that hold negative stereotypes (e.g., women are unlikely to do well on male stereotyped tasks such as mathematical tasks). Gender stereotype boost occurs when people perform tasks with characteristics that are incongruent with gender stereotypes about their social group, namely, tasks that pertain to positive stereotypes about the individual (e.g., men are likely to do well on male stereotyped tasks such as math). In economics, the GSE was demonstrated by Coffman (2014). The author showed that after controlling for ability, women (men) are less likely to put their answers forward as the group lead answers to male (female) stereotyped problems. We expect to observe evidence of GSE in our study. With our experimental design, we can determine whether self-efficacy and belief about one's own ability are key drivers of the GSE.

Gender composition and the activation of GSE

There are a number of ways to make stereotype salient and activate the GSE. Some researchers have used explicit activation by informing subjects directly about stereotypes (cf., Spencer et al., 1999). Others have implemented implicit activation, such as nonconscious priming (e.g., Bargh, Chen, & Burrows, 1996) and identity salience manipulations (Ambady, Shih, Kim, & Pittinsky, 2001). More recently, however, new evidence has cast some doubt on the validity of the above-mentioned activation methods. For example, Doyen et al. (2012) failed to replicate Bargh et al. (1996) and cautioned against the use of nonconscious priming for future research. Flore et al. (2019) did not replicate the findings reported by Ambady et al. (2001) using the identity salience manipulation. The explicit activation design used in Spencer et al. (1999) is likely to suffer from the experimenter demand effect (Zizzo, 2010), resulting confounds in the results. In contrast to the previous literature, we use the gender composition of groups to activate gender stereotypes.

Gender composition of groups can implicitly and subtly activate gender stereotypes, when experimenter demand effect is minimized. Kanter (1977) proposed a theory of tokenism which suggests that the relative number of socially and culturally different people in a group critically shapes a group's interaction dynamics. Notably, the presence of other group members increases the salience of one's group membership and the associated group stereotypes to oneself and to others. In the context of gender, this theory implies that the presence of the opposite gender may activate gender stereotypes. Proceeding in the same spirit, Bordalo, Coffman, Gennaioli, and Shleifer (2016) introduce an economic model of stereotypes based on representativeness heuristics. One key predication of the model is that stereotypes are context dependent and depend on characteristics of a reference group.

Ample empirical evidence supports the view that gender stereotypes can be activated through a group's gender composition. In Cohen and Swim (1995)'s study, individuals in groups that comprised mainly the opposite gender were more likely to report that they expected to be stereotyped by their group members than individuals in groups comprised mostly of the same gender. Sekaquaptewa and Thompson (2003) report that the presence of the opposite gender exacerbates the stereotype threat effect, especially for women. Inzlicht and Ben-Zeev (2000) demonstrate that situational cues, including gender composition, could activate stereotypes and impact individual performance. Hoyt et al. (2010) reveal that the consequences from stereotype threat are more prominent in mixed than single gender groups. In light of this literature, we anticipate our group's gender composition to activate GSE. In particular, we predict that GSE is stronger in mixed gender groups than single gender groups.

Objective performance feedback and the willingness to lead

When the performance quality of an individual is ambiguous and/or difficult to quantify objectively, research in psychology has demonstrated that one is often perceived as less preferred and less competent when performing tasks that are gender-incongruent, thereby resulting in biased performance evaluations (Eagly et al., 1992; Heilman, Wallen, & Fuchs, 2004; Heilman & Haynes, 2005; Heilman & Okimoto, 2007; Heilman & Wallen, 2010; Tosi & Einbender, 1985). Researchers have proposed the prescribed gender roles and stereotype to be a primary source of this effect (Burgess & Borgida, 1999; Eagly & Karau, 2002; Heilman, 2001; Heilman & Haynes, 2005; Rudman & Glick, 2001). However, once objective information on performance is available, gender stereotype no longer influences the performance evaluation (Heilman & Haynes, 2005; Tosi & Einbender, 1985). For example, in an experimental study by Heilman and Haynes (2005), participants worked in mixed gender dyads where individual contributions to a teams' success were ambiguous. The authors found that female leaders were consistently undervalued and viewed as less competent, less effective and less leader-like compared with their male counterparts, unless objective individual performance feedback was given. Given that one is more likely to be rated objectively when individual performance feedback is available, we expect leaders to anticipate this objective assessment and exhibit greater willingness to lead. One thing to note, however, is that much of the above-mentioned research uses tasks that are not incentivized or tasks with only hypothetical consequences. For example, common leadership tasks used in the literature involve artificial scenarios where participants assume a leadership role (see, e.g, Hoyt et al., 2010). The hypothetical nature of the outcome measures brings the reliability into question (Hertwig & Ortmann, 2001). In our study, all decisions are appropriately incentivized, thereby providing a more reliable empirical measurement of behaviors. For example, we measure one's willingness to lead using the position in line the subject selects. We did not frame the decision using an artificial scenario. Instead, we have a fixed rule that implemented a group decision based on answers from those who expressed the strongest desire to lead.

In the economics literature, objective performance feedback has been used in different individual decision domains and shown to be effective in boosting individual performance. For example, Freeman and Gelber (2010) and Azmat and Iriberri (2010) found that the information about both one's own and others' relative skill level helped improve performance. At the same time, there is a large literature on the effect of audience on behavior (see, for example, Andreoni & Bernheim, 2009; Charness, Rigotti, & Rustichini, 2007). In our experiment, with a large group size (audience) and complete objective information about both one's own and other group members' performance, we hypothesize that capable players will demonstrate greater willingness to lead when individual performance feedback information is available publicly.

Experimental design and procedures

The primary goal of our experiment is to test whether the gender composition of a group affects one's willingness to lead through GSE. A second goal is to test whether public performance feedback helps encourage competent players to take the lead, and whether the effect of public feedback differs according to a group's gender composition. We focus on groups that comprised four members. We use a 4×2 between-subject design, where we vary the gender composition (all-male, all-female, majority male (three males and 1 female) or majority female (three females and one male)) and whether performance feedback is public.

We aimed to achieve three goals with our design: 1) to capture the willingness to lead in a setting where groups must make decisions over various gender domains; 2) to vary exogenously the reference group/gender composition; and 3) to vary exogenously whether feedback is public. To accomplish the first goal, we used a modified version of the design reported by Coffman (2014). For the second goal, we implemented a procedure (detailed below) with an eye towards minimizing experimenter demand effects. To accomplish the final goal, we made feedback public and salient using a procedure detailed at the end of this section.

We used ORSEE (Greiner, 2015) to recruit two-hundred-and-forty-eight participants (124 male, 124 female) from a volunteer undergraduate participant pool during May and June 2015. Participants' self-reported ages ranged from 18 to 34 (Mean = 19.79; SD = 1.88); their educational background included a broad range of disciplines, including physical and natural sciences and humanities and social sciences. We conducted a total of 17 sessions, and each session lasted around one and a half hours with an average payment of £17, which was around $38 at the time of the experiment. The participants were paid based on their decisions alone in the experiment, and no show up fee was given.

For each of the experimental sessions, we recruited a gender-balanced sample. We checked subjects in one by one, according to the order in which they arrived. After check-in, each subject was asked to draw a number privately from one of two identical bags. One bag included only odd-numbered balls and the other only even. As we checked in the subjects, the *male* subjects were given the bag with only odd-numbered balls and the *female* subjects drew from the bag with even-numbers. Lab seating was arranged in rows, with each row including four stations. We ensured that the subjects sitting in the same row belonged to the same group, and we told participants that this arrangement was the case. Finally, at each of the stations there was a card with the player's ID.

The experiment was computerized using Ztree (Fischbacher, 2007) and comprised four incentivized parts (Part A, B, C, and D) and a survey that collected demographic information (screenshots of the experiment are included in the appendix). All participants received general instructions informing them that one part of the experiment had been preselected for payment and would be announced at the end of the experiment. They received £1 per point earned on the preselected part. With the exception of the public feedback treatment, participants received no information about their own or others' performance until the experiment concluded. In Appendix A, we provide further details regarding the recruitment process, random group assignment and the feedback mechanism that guaranteed participants' decision anonymity.

Participants faced multiple-choice questions from six categories: arts and literature (Art), entertainment and pop culture (Pop), environmental science (Env), history (Hist), geography (Geo), and sports and games (Sports). Each question included five possible answers and was labeled with its corresponding category (see Fig. 1). Those six categories vary in their perceived gender stereotypeness.

Part A: Individual task

Participants answered 30 multiple-choice questions (5 from each category) on their own. The data from this part provided us with a baseline measurement of individual ability for each category. Subjects received 1 point for each correct answer.

Part B: Willingness to lead in a group

a) Group gender composition revelation
   As the subjects proceeded to part B, they were informed that they would be working with other participants as a group for this part and that they were sitting in the same row as their group members. The experimenter verbally encouraged the subjects to look left and right to observe their group members. Afterwards, participants made decisions about how willing they were to contribute their answers to a *new* set of questions.
b) Willingness to lead – group task
   The subjects made two decisions for each of the new 30 multiple choice questions (see Fig. 2): 1) their answer to the question; and 2) their willingness to lead the group (in other words, put their answer as the group answer by selecting the position in line they would like to stand in the group of four). Since there are four members in the group, there are also four positions in line. Position 1 corresponds to first in line to submit one's own answer as the group answer, position 2 corresponds to second in line to submit one's own answer as the group answer, position 3 corresponds to third in line to submit one's answer as the group answer, and so on.

Among the four group members, the participant who selected the lowest number—the position closest to the front of the line—would have his/her answer submitted as the group's answer. If multiple members selected the same lowest position in line, the computer randomly selected one member's answer as the group answer. The *lower* the position in line, the *more* willing the subject was to lead their group. The payment for this task depended on the submitted group answers. Each group member received 1 point for each correct answer and lost a quarter point for each incorrect answer.

Immediately after subjects checked their group members, and before they started answering a new set of questions (and before public feedback for those in that treatment), they were asked to make incentivized guesses about their own rank within the newly formed group for each of the categories from Part A. They receive additional 25 pence for each correct guess. The purpose of this rank data is to enable insight regarding the effect of group gender composition on self-confidence and to help explain subsequent group task decisions.

Part C and D: Self and group confidence elicitation, risk elicitation

In Part C, we measured participants' confidence in their own answers, as well as in the average answers of their group members. Participants were given the same questions from Part B again, and were asked in an incentive-compatible way (a simplified Becker-DeGroot-Marschak method) to estimate the probability that their own answer was correct and the probability that their group members' answers were correct[2]. Specifically, subjects made three decisions for each question: 1) provide an answer; 2) indicate the probability of their own answer being correct with a number between 1 and 100 – a measure of confidence in one's own answer for question *i*; and 3) provide an estimate

[2] This belief elicitation mechanism is widely used, and its theoretical properties have been studied by Karni (2009), Mobius, Niederle, Niehaus, and Rosenblat (2011) and Schlag, Tremewan, and van der Weele (2015), among others. Under this mechanism, participants are incentivized to provide true beliefs. Appendices A1-A3 provide detailed experimental instructions.

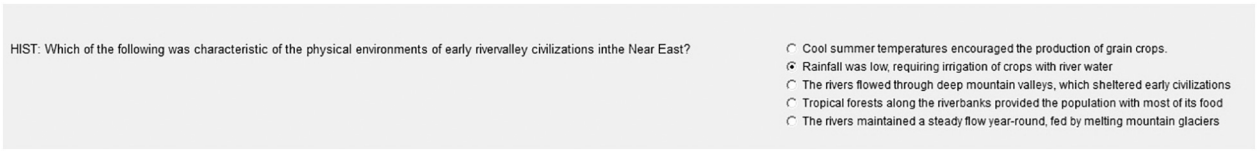

Fig. 1. An example question from Part A of the experiment.

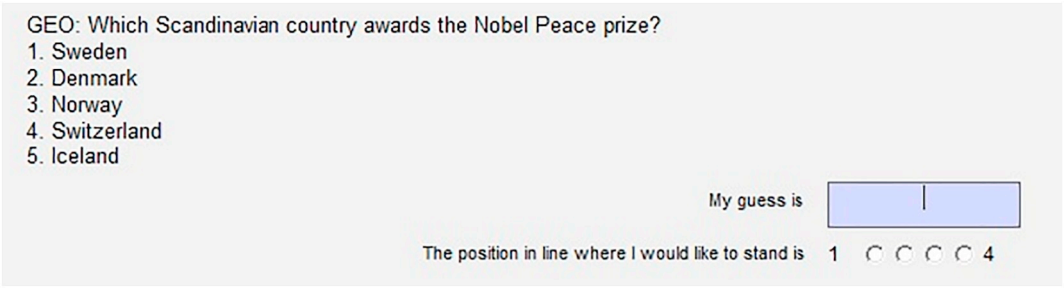

Fig. 2. An example question from Part B of the experiment.

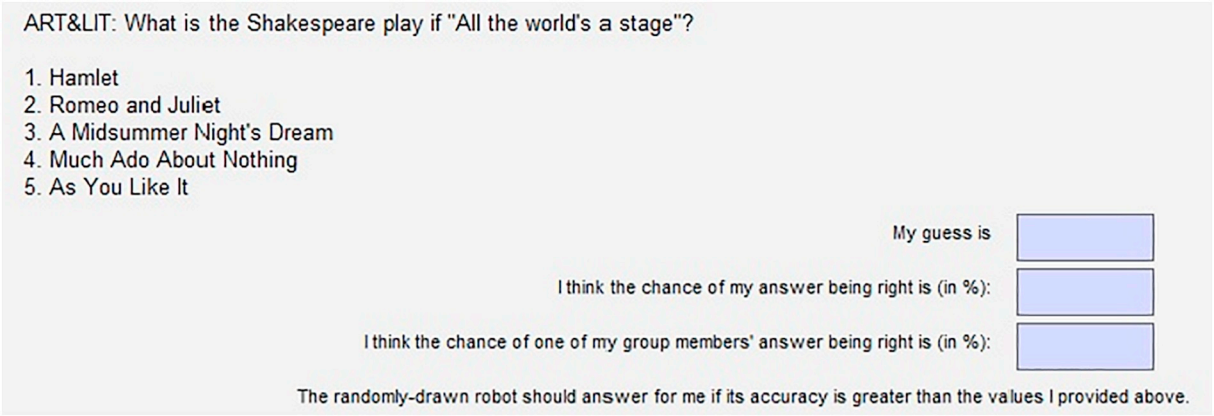

Fig. 3. An example question from Part C of the experiment.

of the probability of the other group members' answer being correct—a measure of confidence in other groups members' answer for question *i*. A correct answer earned half a point, and incorrect answers earned nothing (Fig. 3).

In Part D, we elicited subjects' risk attitudes and asked for demographic and attitudinal information, including, for example, gender, age, where they attended high school, and which question categories they liked or disliked. For this last question, we asked subjects to evaluate the male- or female-typeness of each category. For each of the categories, the subjects were asked to indicate their answers using a slider bar ranging from -1 to 1, where -1 was labeled as "women know more," 1 was labeled "men know more," and the center of the slider bar indicated no gender difference.

Public feedback treatments

Recall that we used a 4×2 between-subject design, varying the gender composition of the group and the availability of public feedback. In the feedback treatments, each participant received an additional page informing her of her Part A performance in each category. In this performance feedback page, the subjects were able to see their own rank, as well as the player ID of the best performer in their group (if the best performer was not them). In the case where the participant happened to be the best performer for the category, instead of her own player ID, the word "You" was boldly displayed in the best performer column. Notice that we informed the subjects about how they and the others in their group performed prior to the group task. The best players knew that their group members also received and acknowledged the fact that they were the best players for the category. Given that player ID was sequential in the group, one could easily know the seating position of the best player in the group.

Knowing the seating position of the group's top performer could seem to impinge on anonymity, potentially opening the possibility that decisions in the lab could have implications for outcomes outside the lab. To the extent that there might be a concern for anonymity, we would argue that this experimental design choice is an ecologically valid feature of our experiment. The reason is that, for most real-life settings, the performance of group members can usually be identified, at least partially. As a result, our design may better approximate the real-world scenario on which we ultimately aim to inform.

That said, as emphasized in Appendix A, it is highly unlikely that any participant's decisions could have implications outside the lab. Reasons are that participants in the lab are generally strangers, and even if friends join the experiment, the chance that they would be randomly assigned to the same group is small. Consequently, participants are unlikely to interact with each other outside the lab. Perhaps more importantly, the public "best performer" information is disconnected from earnings. Recall that earnings are based on one randomly selected part of the experiment, and with only $^1/_4$ chance is this

part related to "Public" information. Further, earnings are based on answers put forward, and the fact that a person was a best performer does not immediately imply that their answers were chosen as the group's answers. Consequently, even in the Public treatment, it is not possible to assign praise or blame for one's earnings to any specific individual.

Predictions

We hypothesize that the gender composition of a group moderates one's willingness to lead a group through GSE, holding ability constant. As a result, some equally (or more highly) capable members may hold back from leading the team and let others take the lead instead. The stepping-back by capable members may result in reduced quality of ideas advanced by the group. Consequently, a group's overall performance can be negatively affected. We detail this hypothesis below.

Conjecture 1. *GSE is stronger in mixed gender groups than single gender groups.*

Following the theory proposed by Kanter (1977) and Bordalo et al. (2016), as well as the literature reviewed in Gender composition and the activation of GSE section, mixed gender groups are likely to activate the gender stereotype, while single gender groups lack the other gender as a comparison group and are thus less likely to activate gender stereotype. Further, one's own gender and the corresponding gender stereotype are more likely to be salient and activated in mixed than in single gender groups. Hence GSE is expected to be stronger for subjects in mixed than single gender groups.

Conjecture 2. *The average quality of group ideas is lower in mixed gender groups than single gender groups. Differences in the quality of contributed group ideas account for group performance differences.*

Following Conjecture 1, women/men in mixed gender groups are more likely to suffer from GSE than those in single gender groups. This conjecture implies that equally capable women or men are more likely to lead teams and offer qualified ideas to *both* male and female stereotyped tasks when they are placed in a single gender group. In contrast, equally capable women (men) may shy away from leading male (female) typed tasks when placed in a mixed gender group. As a result, we expect higher quality ideas from single gender than mixed gender groups. As the submitted group ideas determine group performance, we expect group performance variation can be explained by the quality of a group's ideas.

Conjecture 3. *Public feedback increases the willingness of high-ability players to take the lead.*

In the public feedback treatments, subjects received public information about both their own and other group members' performance. They were able to see their own rank, as well as the player ID associated with the best performer (if the best performer was not self) in her group. In view of the literature reviewed in Objective performance feedback and the willingness to lead section, we hypothesize that the best players are more likely to lead their groups if they are in the public feedback treatments.

Results

Overview and summary statistics

Table 1 below shows the average number of questions answered correctly in Part A (individual task) and B (group task) by gender. For both Part A and B, performance did not differ significantly between men and women for any of the groups with varying gender composition. Overall, men performed significantly better than women. In the analyses that follow, we use the data from Part A to control for general individual ability differences. We also control for whether one answered a specific question correctly in Part B. Additionally, we tested whether the average number of correct answers for a group as a whole changed significantly from Part A to Part B. We do not find statistically significant changes at 5% significance level. It suggests that average group ability did not change due to revelation of group gender composition.

Table 1
Part A & B performance by gender and treatment groups.

Group composition	Number of questions correct		P value ($H_0 : M = W$)	N
	Men	Women		
Part A				
All female		14.22		32
All male	15.50		0.11	32
Female majority	14.78	14.16	0.44	92
Male majority	16.09	15.78	0.71	92
	15.69	14.48	0.003	248
Part B				
All female		14.97		32
All male	16.44		0.07	32
Female majority	15.30	13.77	0.11	92
Male majority	15.70	15.96	0.78	92
	15.81	14.48	0.004	248

Note: P values are from Fisher-Pitman permutation tests for non-binary variables, with a null of equality of distributions between men and women.

Recall that we collected data from our subjects at the end of the experiment (before they received any feedback on their performance) regarding their perception of the gender stereotypeness of each of the categories. The perceived gender stereotypeness did not differ by treatment groups. As a result, we report pooled results in Table 2. Arts & Literature and Entertainment & Pop Culture were considered more female-typed, whereas Sports & Games, Geography and History were regarded as more male-typed. Environmental science was viewed as gender-neutral. Men and women generally agreed on the direction of the stereotypeness of the category. However, they disagreed about the magnitude.

Fig. 4 presents the raw data: average place in line chosen by women and men in Part B by category and treatment. The categories are arranged in increasing order of perceived maleness: Art and Literature (Art) is the least male-typed category and Sports and Game (Spts) the most male-typed category. As the maleness of the category increases, women are less willing to lead the group (the more male-typed the category, the further back women place themselves in line, as revealed by the positive slope of the fitted line), whereas men are more willing to lead (the more male-typed the category, the further up men place themselves in line, as revealed by the negative slope of the fitted line). Both men and women were less likely to vary their positions in line according to the stereotypeness of the category in single gender groups than mixed gender groups. This observation is indicated by a flatter slope of the best-fit line for single than mixed gender groups for both genders. Recall that the position in line was also determined by ability

Table 2
Perceived gender stereotype of categories.

Category	Avg. maleness		Overall Avg.	Normalized z score
	By men	By women		
Art & Literature	−.310	−.386	−.348	−1.189
Entertainment & Pop Culture	−.214	−.253	−.233	−.833
Env. Sci.	.063	−.001	.031	−.007
History	.097	.069	.083	.155
Geography	.137	.051	.095	.191
Sports & Games	.612	.532	.573	1.683

Note: The elicitation is on the scale of −1 (female knows more) to 1 (male knows more). The more positive the number, the more male-stereotyped the category, and more negative indicates more female.

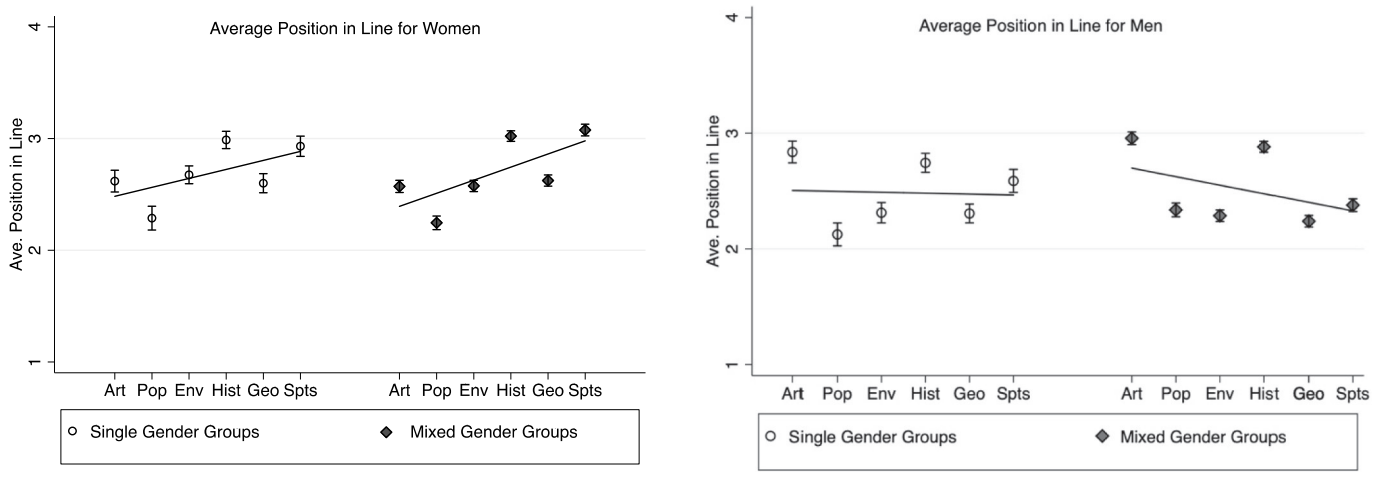

Fig. 4. Average position in line by gender and treatment. Note: Error bar = mean ± standard error of the mean. The category is ranked by increasing perceived maleness of the category. Lower position number indicates a greater willingness to lead the group.

(described by Table 1) as well as one's perceived gender stereotypeness of the category (described by Table 2). We control for those factors in the next section.

Main results

Evidence of gender stereotype effect

Table 3 Regression (1) reports the first evidence on GSE. We regress a participant's chosen position in line for question *i* on gender (*Female Dummy*), the z-score of reported "maleness" of the category from which question *i* is drawn (*Maleness of Category*), the interaction of gender and "maleness" score (*Female* x *Maleness*). For robustness checks, we also include a standard set of controls in Regression (2). Some of the controls are used throughout the analyses reported in this paper: whether or not one answered question *i* correctly (*Answered Qn. i Correctly* – a proxy for her question-specific ability) and her Part A score in the category from which question *i* was drawn (*Part A Score* – a proxy for her broader ability in that specific category), dummies for the treatments, race dummies, a dummy for attending secondary school in the UK, and the overall probability of a correct answer for question *i*[3] in Part B. Errors are clustered at the individual level. Because the dependent variable is the position in line, lower coefficient estimates indicate a greater willingness to contribute.

We find that as the maleness of the category increased, men became significantly *more* likely to lead the group (demonstrated by the significantly negative coefficient of *Maleness of Category*, $p < .01$), when holding ability constant. Women, in comparison, became significantly *less* likely to lead as the maleness of the category increased (shown by the significantly positive sum of the coefficients of *Maleness of Category* and *Female* x *Maleness*, $p < .01$). Our results are qualitatively similar to those of Coffman (2014); however, the size of the effect in our data is more than twice the level she reports.

Evidence of gender composition moderating GSE

The analysis in Table 3 (1) and (2) above establishes that men respond to increased *maleness* of the category with increased leadership, and women respond to increased *femaleness* of the category (or decreased *maleness* of the category) by doing the same. That is, both men and women show GSE when pooling all treatments, and holding ability constant. We now turn to our key question: does the gender composition of a group moderate the observed *GSE*? The answer is yes.

[3] This is calculated as the percentage of subjects who answer the question *i* correctly. This variable controls the overall difficulty of the question.

In Table 3 Regression (3), we include additional regressors: the interaction of the maleness score and a dummy for mixed gender treatment (*Maleness* x *Mixed Gender*) and the interaction of the gender dummy, the maleness score and the mixed gender group dummy (*Female* x *Maleness* x *Mixed Gender*). It is clear that men in an all-male group do not exhibit GSE: their willingness to lead does not differ by the maleness of the category (demonstrated by the statistically insignificant coefficient of *Maleness of Category*, $p = 0.22$). In contrast, men do display GSE in mixed gender groups as shown by the significantly positive sum of the coefficients of *Maleness of Category* and *Maleness* x *Mixed Gender*, $p < .01$. The effect in mixed gender groups is not only highly significant, but the size is also large in magnitude. Unlike men, women exhibit GSE even in women-only groups: women are less likely to lead when the *maleness of the category* increases (shown by the significantly positive sum of the coefficients of *Maleness of Category* and *Female* x *Maleness*, $p < .01$). Moreover, the size of GSE almost doubles, when women are placed in a mixed gender group (shown by the significant coefficient of *Female* x *Maleness* x *Mixed Gender*, $p < .01$). With additional controls in Table 3 Regression (4), our results hold. Further

Table 3
OLS predicting position in line for Question i in Part B - Willingness to lead.

	(1)	(2)	(3)	(4)
Female Dummy	0.152***	0.057	0.150***	0.055
	(0.05)	(0.04)	(0.05)	(0.04)
Maleness of category	−0.403***	−0.368***	−0.165	−0.169
	(0.07)	(0.05)	(0.13)	(0.10)
Female x Maleness	0.832***	0.709***	0.480***	0.368***
	(0.09)	(0.08)	(0.17)	(0.14)
Maleness x Mixed gender			−0.327**	−0.273**
			(0.15)	(0.12)
Female x Maleness x Mixed gender			0.493**	0.480***
			(0.20)	(0.17)
Constant	2.533***	4.128***	2.533***	4.126***
	(0.04)	(0.13)	(0.04)	(0.13)
Controls	No	Yes	No	Yes
Observations	7440	7440	7440	7440
R^2	0.028	0.309	0.030	0.310

Notes: Lower position in line indicated greater willingness to lead. The unit of observation is question i. Each participant in the experiment answered 30 questions. Cluster-robust standard errors at individual level were used in the regressions (248 clusters in total). Controls include a dummy for whether question i was answered correctly, Part A score for the category, race dummy, UK secondary school dummy, and the overall probability of a correct answer for question *i* in Part B.
** Indicates significance level at 10%, ** at 5%, *** at 1%.*

Table 4
OLS predicting Part A Belief in group ranking - Impact of group composition and effect of stereotypes.

	Single gender (1)	Mixed gender (2)	Pooled (3)
Female Dummy	0.160	0.230***	0.223***
	(0.11)	(0.07)	(0.06)
Stereotypeness	−0.276**	−0.598***	−0.272**
	(0.14)	(0.07)	(0.14)
Stereotypeness x Mixed gender group			−0.323**
			(0.15)
Part A Score - Same category	−0.121***	−0.104***	−0.109***
	(0.04)	(0.02)	(0.02)
Constant	2.984***	2.603***	2.830***
	(0.20)	(0.10)	(0.10)
Controls	Yes	Yes	Yes
Observations	384	1104	1488
R^2	0.089	0.152	0.136

Notes: Lower number in ranking indicated greater confidence. The unit of observation is category *i*. Each participant in the experiment reported ranking belief for 6 categories. Cluster-robust standard errors at individual level were used in the regressions. Controls include race dummy, UK secondary school dummy. Standard errors are clustered at individual level.
** Indicates significance level at 10%, ** at 5%, *** at 1%.*

investigation indicates that the results reported above are also not driven by any of the mixed gender groups. In fact, GSE occurs with even one member of the opposite gender present (see Appendix Tables A1 and A2).

In sum, we conclude that the presence of the opposite gender significantly activates GSE, while the absence of the opposite gender in a group mitigates this effect. Further, men seem not to display GSE when women are absent. An implication is that we should observe higher overall percentages of 1st in line answers in single as opposed to mixed gender groups. The reason is that as GSE dials down for single gender groups, equally capable players from both genders are equally likely to lead in all gendered domains. As we discuss further in later sections, the percentage of 1st in line answers is critical for group performance.

Gender composition moderates the GSE – mechanism through beliefs

In this section, we provide evidence that gender composition of a group may change one's belief about her own standing in the group, and the changes in beliefs in turn impact GSE. Recall that immediately after the random group assignment and prior to Part B, we asked our participants to guess their ranking in the newly formed group for each of the six categories from Part A.

In Table 4, we show that controlling for Part A performance (own ability), people perceived their own group ranking very differently depending on the gender composition of their randomly assigned group. We regress one's perceived group rank for category *i* with a gender dummy (*Female Dummy*), the absolute value of one's reported "maleness" z-score of the category *i (Stereotypeness)*, the interaction of *Stereotypeness* and a mixed gender group dummy, the number of questions answered correctly for category *i* and other standard controls reported in the previous analyses. The coefficient of *Stereotypeness* measures how one's perception of her group ranking for category *i* changes according to the level of *gender-congruence* of that category. The negative sign of the coefficient means that the more gender-congruent the category is, the higher one ranks herself in the group. The more negative the coefficient, the greater the effect the stereotypes have on her belief about her standing in the group.

Consistent with the results from the previous section, we find that, holding ability constant, beliefs systematically vary according to the gender-congruence of the category, and that this effect is greater in mixed than single gender groups. Indeed, men in single gender groups did not vary their group rank beliefs with the stereotypeness of the category (for details see appendix Table A3). Further, group gender composition manipulation does not impact beliefs over average ability of other group members (see appendix, Table A4).

Table 5
Part B Group performance by Group composition.

Group composition	Avg. performance (in points)	% of Answers from 1st in line	N
All male	19.69	63%	8
	(0.77)		
Male majority	18.04	61%	23
	(0.82)		
All female	16.09	50%	8
	(1.12)		
Female majority	15.54	56%	23
	(0.72)		

Note: Standard errors are reported in parentheses.

Overall group performance analyses

We next consider overall group performance across different treatments. We demonstrated in the previous section that a group's gender composition moderates people's willingness to lead, particularly in a gender-incongruent domain. People in single gender groups are more likely to lead in all areas than mixed gender groups, holding ability constant. As a result, groups with different gender compositions may have different fractions of 1st in line answers contributed as group answers. Moreover, the fraction of 1st in line answers may help to explain differences in overall group performance. Table 5 below summarizes performance results by treatment. All-Male and Majority-Male groups performed significantly better than Majority-Female groups ($p < .01$). About 62% of the submitted group answers for All-Male and Majority-Male groups were from the 1st in line, whereas only 53% of the group answers in All-Female and Majority-Female groups were from 1st in line answers[4].

As shown in the regression analyses in Table 6, the fraction of 1st in line answers is a highly significant predictor for group performances (Table 6, Regression 1), even after controlling for the average ability of the group, which we denote as *Group Part B Score* (Table 6, Regression 2). Further, the group performance differences (as shown in Table 6, Regression 4) disappear when we control for the percentage of answers from 1st in line and average group ability (Table 6, Regression 3).

The percentage of answers from 1st in line were of great importance, as detailed in Table 7. Answers from 1st in line were about 89% accurate, but the accuracy rate drops to 57% if the answers are from 2nd in line. Given the position in line, we do not find gender differences with regard to the rate of accuracy. However, Table 8 shows that, conditional on having correct answers, men were significantly more likely than women to choose to lead. The implication is that capable female players were not realizing their full potential by leading the group. We also observe that conditional on an incorrect answer, men were more likely to try to lead by placing themselves at least third in line. Doing this, however, had little negative impact on group performance. The reason is that it was rarely the case that answers from 3rd in line were used as the group answer.

We now turn to the public feedback conditions in order to investigate whether this encourages high ability players, and especially high ability women, to choose to lead.

Effect of public feedback on the best players

Recall that in our public feedback treatment we provided players with information about their own rank, as well as the ID of the best player in their group. In Table 9, we regressed participants' chosen

[4] For all groups, around 30% of the answers were from 2nd in line. A very small fraction of the answers was from 3rd or 4th in line.

Table 6
OLS on group performance.

	(1)	(2)	(3)	(4)
Fraction of 1st in line answers	15.761***	8.038***	7.707**	
	(3.59)	(2.39)	(2.89)	
Group Part B score		1.353***	1.301***	
		(0.12)	(0.13)	
All female			0.181	0.550
			(1.11)	(1.30)
All male			0.835	4.144***
			(0.93)	(1.04)
Male majority			0.444	2.500**
			(0.72)	(1.10)
Constant	7.967***	−8.067***	−7.390***	15.543***
	(2.07)	(1.83)	(1.99)	(0.728)
Observations	62	62	62	62
R^2	0.225	0.668	0.673	0.163

Notes: The unit of observation is group *i*. There were 62 groups in total in the experiment. Robust standard errors are reported in parentheses. The control group is the female majority group.
** Indicates significance level at 10%, ** at 5%, *** at 1%.*

Table 7
Part B Accuracy by position in line.

Position in line	Answer accuracy rate		P value ($H_0 : M = W$)	N
	Men	Women		
1st In line	89%	88%	0.84	1934
2nd In line	57%	54%	0.39	1405
3rd In line	40%	36%	0.09	1836
4th In line	24%	27%	0.13	2265

Notes: P values are from regressions with accuracy rate as the dependent variable and gender dummy as the independent variable. The unit of observation is question *i*. Cluster-robust standard errors at the individual level were used.

Table 8
Part B Average position in line by gender.

Average position in line	Men	Women	P value ($H_0 : M = W$)	N
Correct answers	1.95	2.13	0.00	3757
	(0.04)	(0.05)		
Incorrect answers	3.12	3.21	0.16	3683
	(0.04)	(0.06)		

Notes: P values are from regressions with position in line as the dependent variable and gender dummy as the independent variable. The unit of observation is question *i*. Cluster-robust standard errors at the individual levels are reported in parentheses.

position in line for question *i* with the regressors used in previous analyses, and an additional set of *Feedback* regressors. Here, *Feedback* is a dummy indicating whether a participant is in public feedback treatment. *Female* x *Feedback* is an interaction term that measures whether the effect of feedback on women is different than the effect on men. As a result, the coefficient of *Feedback* indicates the effect of feedback on men, and the sum of *Feedback* and *Female* x *Feedback* represents the total effect of feedback on women. Overall, we find strong evidence that public feedback encourages the best *female* players to take the lead (F test for H_0: Feedback + Female x Feedback = 0, $p < .01$), as shown in Table 12 (1) pooled analyses. We also included interaction terms regarding feedback and the maleness of the category. None of those interaction terms were statistically significant. Detailed tables are included in the appendix, Table A4.

It is interesting to note that the effect of feedback depends on the gender composition of the group. In single gender groups, the high-ability men and women were both significantly affected by public feedback. On the one hand, the best male players in an all-male group responded significantly positively to feedback by leading more. Indeed, they moved up in the line by about 12%[5]. On the other hand, the best female players were deterred by public feedback and responded by taking a step back (possible explanations for this finding are offered in the following section). In mixed gender groups, the best male players did not seem to be affected by the feedback, while the best female players responded positively by leading the group more ($p < .01$ for both Majority-Female and Majority-Male groups). We did not observe an interaction effect between feedback and the gender stereotype of the category, in other words, the effect of the feedback did not differ by the maleness of the category. Detailed regressions are reported in the appendix, Table A4.

Discussion

In this paper, we find that a group's gender composition significantly moderates GSE. In particular, participating in a mixed gender group (even as the majority gender) substantially increases the impact of GSE, while being in a single-gender group diminishes (and for men eliminates) this effect. Consequently, capable members of groups with mixed gender compositions choose whether to lead and contribute differently. Moreover, we show that group performance differences can be largely explained by the fraction of capable players who choose to lead. Additionally, we find that public revelation of objective performance increases the chance that men in all-male groups will prefer to take the lead; surprisingly, however, this public revelation has the opposite effect for women in all-female groups—capable women are deterred from leading under public revelation. In mixed gender groups, however, public feedback significantly encourages the best female players to lead. So far, we have left open the possible channels that the presence of the opposite gender may activate GSE. In the next subsection, we discuss the possible channels and the existing evidence.

Possible channels for activation of GSE

There are two channels through which the presence of the opposite gender may activate GSE. If women believe that their male team members are relatively better at male-typed tasks and worse at female-typed tasks holding ability constant, then they will choose to step back in male-typed tasks and step up in female-typed tasks. We refer to this channel as the gender comparative advantage channel. It follows that the presence of the opposite gender activates GSE, since there is a group with comparative advantage. We anticipate GSE to *disappear* in single gender groups (no one in the group has a particular comparative advantage) and *reappear* in mixed gender groups. Alternatively, if women simply believe they are less capable at male-typed tasks per se, then they will step back in male-typed tasks, even when there is no male presence. We call this channel the gender identity channel. Under the identity channel, we predict that GSE can impact behavior even in single gender groups. Moreover, because gender identity is salient in mixed-gender groups, under this channel GSE is stronger in mixed than single gender groups.

We find that men suffer from GSE in mixed but not single gender groups, whereas women experience GSE in both types of groups. This finding suggests that GSE is more likely to operate through the gender comparative advantage channel for men, but through the gender identify channel for women.

Discussion of the results and implications

Gender diversity has been the focus of many public-policy debates,

[5] The overall average position in line is about 2.5, the increase of the position in line for best male players in all-male groups is 0.304, about a 12% increase.

Table 9
OLS predicting position in line for Question i in Part B - Impact of feedback for players with best Part B score in category.

	Pooled (1)	All female (2)	Female maj. (3)	Male maj. (4)	All male (5)
Female Dummy	0.115**		0.288***	0.172*	
	(0.05)		(0.08)	(0.10)	
Maleness of category	−0.370***	0.389***	−0.551***	−0.455***	−0.068
	(0.04)	(0.08)	(0.11)	(0.06)	(0.09)
Female x Maleness	0.779***		1.007***	0.812***	
	(0.06)		(0.13)	(0.12)	
Feedback	−0.068	0.192**	−0.012	0.023	−0.315***
	(0.04)	(0.10)	(0.10)	(0.06)	(0.09)
Female x Feedback	−0.041		−0.109	−0.235**	
	(0.06)		(0.11)	(0.12)	
Answered Qn. i Correctly	−0.784***	−0.685***	−0.735***	−0.772***	−0.990***
	(0.03)	(0.09)	(0.05)	(0.05)	(0.09)
Part A Score - Same category	−0.036***	−0.119***	−0.021**	−0.032***	−0.065***
	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)
Constant	4.285***	5.225***	3.852***	4.029***	4.775***
	(0.11)	(0.38)	(0.18)	(0.19)	(0.24)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	4950	510	1890	1920	630
R^2	0.305	0.392	0.302	0.292	0.389

Notes: Lower position in line indicated greater willingness to lead. The unit of observation is question i. Each participant in the experiment answered 30 questions. Cluster-robust standard errors at individual level were used in the regressions (248 clusters in total). Controls include treatment dummies, race dummy, UK secondary school dummy, the overall probability of a correct answer for question *i* in Part B.
Indicates significance level at 10%, ** at 5%, * at 1%.*

with special attention paid to gender diversity in the high-tech industry. Yet it is far from clear how gender diversity impacts group economic performance, and through which channels it operates. We move towards answering this question by exploring whether gender composition may affect group performance by impacting the willingness to lead of those most capable.

Using groups of four, we observed that both men and women are less likely to take the lead on problems outside of their own gender-stereotyped domain. Further, we found that a group's gender composition moderates this effect. Specifically, both women and men placed in single gender groups were at least 50% less likely to experience the *gender stereotype effect* than when placed in mixed gender groups. While *GSE* vanished for men in all-male groups, women continued to display this effect even when placed in all-female groups (though it was substantially mitigated). A crucial implication is that GSE may operate through different channels for men and women, particularly the channels of gender comparative advantage for men, and gender identity for women.

We observed that GSE can be explained by changes in beliefs. For example, we found that a woman 's belief about her ability ranking within a group is dramatically impacted by a group's gender composition. Importantly, the direction of her change in beliefs is consistent with the impact of GSE. One reason that women display GSE even in all-female groups may be that gender identity is deeply rooted for women, and the presence of a man may not be needed to remind a woman of her femaleness. There is much evidence for this finding, including, e.g., that females are more susceptible than males to gender-stereotyped prescriptions of appropriate social behavior (Burgess & Borgida, 1999; Heilman et al., 2004; Rudman & Glick, 2001). As a result, special attention should be paid to female leaders, since women may be more susceptible to gender stereotype threats than men (Kiefer & Sekaquaptewa, 2007). Indeed, Karpowitz, Monson, and Preece (2017) demonstrate that a simple verbal message intervention from party leaders can significantly encourage women to run and ultimately win positions as precinct leaders.

A closer look at overall group performance reveals that the key to group success is to have more answers from the 1st in line (i.e., for capable players to lead the group), as 1st in line answers have the highest accuracy rate. Moreover, the fraction of 1st in line answers is influenced by the gender composition of the group. We demonstrated that gender composition moderates people's willingness to lead the groups and further influence the overall group performances. We also found that conditional on a correct answer, men were significantly more likely than women to take the lead. The implication is that there are missed opportunities from capable female players. Consequently, we investigated whether it might improve the efficiency of group decision-making to provide public feedback to participants by providing not only their own group rank (relative performance), but also the ID of the best player in their group. Overall, this intervention successfully resulted in greater numbers of high-ability female leaders.

Further, we found the effect of public feedback to vary according to the gender composition of the group. In single gender groups, the best male players responded to positive feedback by leading more, whereas the best female players seemed to be deterred from taking control of the group. One explanation for this observation could be that women care more about fairness and would like to signal their cooperativeness by letting others shine as well (see, e.g., Andreoni & Vesterlund, 2001; Charness & Rustichini, 2011). Alternatively, women in all-female groups may believe that promoting themselves, and outshining all other women, could lead her to be shunned by other group members (Rudman, 1998). Gneezy and Rustichini (2004) offer evidence to support this finding. They find girls in all-female racing groups performed worse than girls in mixed gender competitions. In a similar spirit, we found that the best female players responded to positive feedback by taking the lead more in mixed gender groups.

Our results connect to the findings of Babcock, Recalde, Vesterlund, and Weingart (2017). Those authors show that women in gender-diverse environments are more likely than men to accept jobs with low probabilities of promotion. In particular, they find that in single gender groups, men and women are equally likely to volunteer, but only in

mixed gender groups do women volunteer more than men. This behavior is consistent with our findings to the extent that volunteering is a female-stereotyped domain. Our results are also in line with Born, Ranehill, and Sandberg (2018). Born and coauthors find that women are less willing to lead in teams that mainly comprised of men. Finally, our data provide support for institutional policies that encourage women and men to lead and not to shy away from success, especially in gender-incongruent fields.

Our findings have especially important implications for team formation in gendered industries (e.g., the tech sector) as compared to more traditionally gender-neutral industries (such as Media and Entertainment). We find GSE to be exacerbated by the presence of even one person from the opposite gender; thus, women who work in gendered industries should be especially encouraged to avoid shying away from leadership opportunities, as doing so may result in their being overlooked for promotion or advancement opportunities.

Furthermore, as numerous studies have shown, ambiguous performance metrics lead to biased performance evaluations, particularly for women performing male-typed tasks (see, e.g., Heilman et al., 2004; Heilman & Haynes, 2005). Biased evaluations may in turn lead to fewer women pursuing traditionally male roles, and particularly leadership roles. The reason is that the quality of leadership is difficult to measure, and women may in turn expect to receive disproportionately less credit for success and generally less favorable performance evaluations. However, our study suggests that public performance feedback may work to address this concern. Examples of approaches organizations might pursue include publicizing numbers of sales and corresponding revenues, the numbers of projects successfully completed, or even, in academics, the numbers and placements of papers published or grants awarded. It seems simple to implement such policies, and our results suggest that doing so may mitigate the GSE and help encourage the most capable women to choose to lead.

Limitations and future research

Although there are many unique and important aspects to our findings, there is little question that the methodology we used in our research limits the degree to which we can extrapolate to natural environments. Our study used undergraduates as participants and, although the majority of them had work experience, the type of work experience may be limited and correspondingly limit our ability to generalize from our data. Future research using different participant pools, such as MBA students, may offer additional validation and insight for our findings. Further, the current experimental group work paradigm is suitable for the university student population, as multiple-choice questions are common tasks. However, this paradigm may not be appropriate for other participant pools, such as employees in large organizations. Future research with these participants may require corresponding design adaptations. Finally, it would be beneficial to study the willingness to lead and test the effect of public performance feedback in a natural field setting using non-experimental methods. Willingness to lead could be proxied with the data on people who ask for a promotion. One form of public feedback could be the employee recognition awards many companies implement. One could test whether the recognition award increases women's probability of subsequently asking for a promotion. Given the endogeneity issue associated with the data, one might wish to apply the regression discontinuity design by comparing employees with a set of similar characteristics who are runners-up for employee recognition awards.

We find that public performance feedback greatly encourages capable women to step up and take the lead, even for tasks that are stereotyped as existing in the male domain. While we recognize that objective and public feedback may be difficult to implement in practice, especially given that objective performance may not be readily available in some environments that require leadership, alternative public feedback such as achievements in training or certification may be provided as evidence of competence for high ability individuals. Our study mainly addresses the instrumental leadership where organizations rely on the leader's expertise. Whether our results can extend to other forms of leadership is still an open question for future research. Finally, we focus on gender stereotypes as the main mediator for one's willingness to lead. It is possible that mate selection as posited by evolution theorist may offer some alternative explanations to some of our observed gender differences, because different strategies would have benefitted men versus women in our distant ancestral past (cf., Davies & Shackelford, 2008). Future research may be able to address this possibility.

Conclusion

The value and importance of gender diversity in organizations is well-understood. Nonetheless, public policy debates continue regarding how best to achieve this diversity. These debates can be informed, and policy advice strengthened, by improving our knowledge about the channels through which diversity impacts economic outcomes. This paper is an effort to address this issue. We discovered that, through the gender stereotype effect, gender composition affects group performance by impacting the most capable members' willingness to lead. Importantly, we found that both genders are more likely to experience the gender stereotype effect and shy away from leadership in gender-incongruent fields when the workplace is gender diverse. Single-gendered workplaces are not a solution, as women, in particular, continue to suffer from GSE even in the absence of men. Our evidence suggests that the most capable female and male leaders emerge, and consequently the best group outcomes are obtained, when public performance feedback is provided to mixed-gender groups. This policy is both highly beneficial and often straightforward to implement, meaning it should be of great value to any economic or social organization. It is also worth cautioning that there is greater heterogeneity in tasks people perform in certain industries, thereby making it challenging to make comparisons.

Appendix A. Steps taken to help ensure participants' anonymity

A.1. Recruitment details for experimental sessions

Our ORSEE recruitment system includes over 1500 registered participants. The pool is refreshed at the start of every academic year to minimize the number of the inactive subjects. All accounts are made inactive at the start of the academic year, and subjects are required to reactivate their accounts themselves. All incoming students also receive invitations to join the pool. The pool includes students from 35 different disciplines, over 200 degree programs and spans across four stages of study.

For each experimental session, invitation emails were sent to a randomly selected subset of the subject pool. Only those who accepted the session invitation were able to participate in any particular experimental session. For the experiments reported above, we only used 248 out of over 1500 potential subjects (less than 20% of the total available subjects' pool). Therefore, given the 1500 subjects in the recruiting pool, it is very unlikely to have subjects from the same discipline, same program and year (those subjects are mostly likely to know each other outside of the laboratory) in the same experimental session.

A.2. Random assignment of the groups

Our experimental sessions included 16 subjects in groups of four, and at the start of the experiment, every subject drew a random number which determined their group. Consequently, even in cases where some participants were in the same discipline, same program, same year and might know each other, the probability of those subjects being put into the same group was only about 25%.

A.3. Feedback and payment mechanism

Subjects were given very limited feedback during the experiment. They did not know about their overall group performance in the majority of the cases. There were four incentivized parts in the experiment, but only one of those parts was randomly selected for payment. The subjects did not receive any feedback during the experiment (with the exception of the public feedback treatment where the subjects were only informed how well each group member performed in the experiment's previous part). At the end of the experiment, subjects were informed about which part was chosen for payment and the amount they earned for that part. No feedback was given for parts not selected for payment. For example, if Part A was selected for payment, the subjects would be paid according to their individual performance in Part A, and they would be able to infer how well they did in Part A. However, they would not learn how well they or their group did in Parts B, C or D. As a result, only around 25% of the groups might have learned how well their group performed.

Even in those cases where they were able to learn their group's performance, there was no information that could be used by participants to attribute praise or blame to any specific group member. Note that the overall group performance is determined by the **submitted** group answers. The subjects were not informed of whose answers were submitted as group answers. As a result, it was impossible for any subject to assign praise or blame to any specific person, and thus they were unable to reward or punish outside (or inside) of the laboratory.

Table A1
OLS predicting Part A Belief in group ranking - Impact of group composition on male.

	All male (1)	Majority male (2)	Minority male (3)
Maleness of category	−0.175	−0.540***	−1.216***
	(0.20)	(0.09)	(0.20)
Part A Score in category	−0.087	−0.092**	−0.047
	(0.06)	(0.04)	(0.07)
Constant	3.020***	2.736***	2.524***
	(0.33)	(0.14)	(0.28)
Controls	Yes	Yes	Yes
Observations	192	414	138
R^2	0.069	0.143	0.260

Note: Lower number in ranking indicated greater confidence. Standard errors are clustered at individual level. Controls include race dummy, UK secondary school dummy.
** Indicates significance level at 10%, ** at 5%, *** at 1%.*

Table A2
OLS predicting Part A Belief in group ranking - Impact of group composition on female.

	All female (1)	Majority female (2)	Minority fFemale (3)
Femaleness of category	−0.369*	−0.584***	−0.430**
	(0.20)	(0.11)	(0.18)
Part A Score in category	−0.154***	−0.099***	−0.167***
	(0.05)	(0.03)	(0.05)
Constant	3.161***	2.687***	3.034***
	(0.22)	(0.16)	(0.23)
Controls	Yes	Yes	Yes
Observations	192	414	138
R^2	0.109	0.116	0.169

Note: Lower number in ranking indicated greater confidence. Standard errors are clustered at individual level. Controls include race dummy, UK high school dummy.
Indicates significance level at 10%, ** at 5%, * at 1%.*

Table A3
OLS predicting Part B group confidence - (No) Impact of group composition.

	For female (1)	For male (2)
Female majority	−3.134	−2.997
	(1.97)	(2.66)
Male majority	−0.000	−1.768
	(2.80)	(1.98)
Self confidence	0.439***	0.427***
	(0.03)	(0.02)
Constant	33.245***	28.790***
	(3.20)	(2.73)
Controls	Yes	Yes
Observations	3720	3720
R^2	0.450	0.446

Note: Standard errors are clustered at individual level. Controls include race dummy, UK secondary school dummy, the overall probability of a correct answer for question *i* in Part B.
** Indicates significance level at 10%, ** at 5%, *** at 1%.*

Table A4
OLS predicting position in line for Question i in Part B - Impact of feedback for players with best Part B score in category.

	Pooled (1)	All female (2)	Female maj. (3)	Male maj. (4)	All male (5)
Female Dummy	0.123**		0.286***	0.163*	
	(0.05)		(0.08)	(0.10)	
Maleness of category	−0.291***	0.421***	−0.612***	−0.381***	0.033
	(0.08)	(0.16)	(0.16)	(0.12)	(0.13)
Female x Maleness	0.626***		0.842***	0.897***	
	(0.11)		(0.19)	(0.19)	
Feedback	−0.060	0.191**	−0.021	0.030	−0.304***
	(0.04)	(0.10)	(0.10)	(0.06)	(0.09)
Female x Feedback	−0.050		−0.104	−0.227**	
	(0.06)		(0.11)	(0.12)	
Feedback x Maleness	−0.119	−0.043	0.115	−0.096	−0.194
	(0.09)	(0.18)	(0.22)	(0.14)	(0.17)
Female x Feedback x Maleness	0.240*		0.284	−0.200	
	(0.13)		(0.26)	(0.25)	
Answered Qn. i Correctly	−0.783***	−0.687***	−0.733***	−0.777***	−0.988***
	(0.03)	(0.09)	(0.05)	(0.05)	(0.09)
Part A Score	−0.036***	−0.119***	−0.021**	−0.032***	−0.066***
	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)
Constant	4.282***	5.225***	3.860***	4.030***	4.786***
	(0.11)	(0.38)	(0.18)	(0.19)	(0.24)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	4950	510	1890	1920	630
R^2	0.305	0.392	0.305	0.2924	0.388

Note: Lower position in line indicated greater willingness to lead. Standard errors are clustered at individual level. Controls include treatment dummies, race dummy, UK secondary school dummy, the overall probability of a correct answer for question *i* in Part B.
** Indicates significance level at 10%, ** at 5%, *** at 1%.*

Table A5
Variable means, standard deviations and correlations.

	Variable	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13
1	Position in line	2.60	1.17													
2	Female	0.50	0.50	0.08												
3	Maleness	0.03	0.42	−0.001	−0.08											
4	Female x Maleness	0.001	0.30	0.11	0.003	0.70										
5	Maleness x Mixed gender	0.03	0.36	−0.01	−0.06	0.84	0.57									
6	Maleness x Mixed gender x Female	0.004	0.25	0.10	0.02	0.58	0.83	0.69								
7	Answered Q *i* Correctly	0.50	0.50	−0.48	−0.04	−0.01	−0.05	0.01	−0.03							
8	Part A Score	15.08	3.23	−0.19	−0.19	0.05	0.02	0.05	0.02	0.14						
9	Asian	0.17	0.37	0.13	0.07	−0.03	−0.01	−0.02	0.01	−0.12	−0.43					
10	Mixed race	0.06	0.23	−0.02	0.10	−0.01	0.002	0.002	0.01	0.01	0.06	−0.11				
11	Other race	0.01	0.11	0.03	0.11	−0.002	0.01	−0.01	0.00	−0.02	−0.11	−0.05	−0.03			
12	UK secondary school	0.71	0.46	−0.16	−0.13	0.02	−0.01	0.01	−0.02	0.11	0.49	−0.47	0.08	−0.01		
13	Q *i* Difficulty	0.50	0.24	−0.40	0.00	−0.04	−0.03	−0.04	−0.03	0.48	−0.00	0.00	−0.00	−0.00	0.00	

Fig. A1. Part C Instruction I.

Fig. A2. Part C Instruction II.

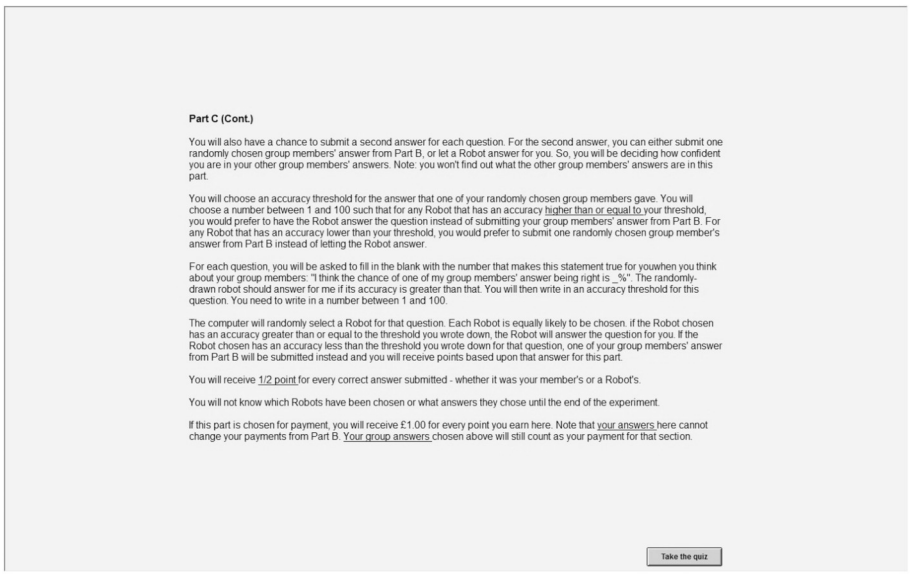

Fig. A3. Part C Instruction III.

References

Adams, R. B., Funk, P., Barber, B., Ho, T., & Odean, T. (2012). Beyond the glass ceiling: Does gender matter? *Management Science, 58*(2), 219–235. https://doi.org/10.1287/mnsc.1110.1452.

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science, 12*(5), 385—390.. https://doi.org/10.1111/1467-9280.00371.

Anderson, M. H., & Sun, P. Y. T. (2017). Reviewing leadership styles: Overlaps and the need for a new 'full-range' theory. *International Journal of Management Reviews, 19*(1), 76–96. https://doi.org/10.1111/ijmr.12082.

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica, 77*(5), 1607–1636. https://doi.org/10.3982/ECTA7384.

Andreoni, J., & Vesterlund, L. (2001). Which is the fair sex? Gender differences in altruism. *Quarterly Journal of Economics, 116*(1), 293–312. https://doi.org/10.1162/003355301556419.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *Leadership Quarterly*. https://doi.org/10.1016/j.leaqua.2010.10.010.

Antonakis, J., & House, R. J. (2014). Instrumental leadership: Measurement and extension of transformational-transactional leadership theory. *Leadership Quarterly, 25*(4), 746–771. https://doi.org/10.1016/j.leaqua.2014.04.005.

Apesteguia, J., Azmat, G., & Iriberri, N. (2012). The impact of gender composition on team performance and decision making: Evidence from the field. *Management Science, 58*(1), 78–93. https://doi.org/10.1287/mnsc.1110.1348.

Aronson, J., Quinn, D. M., & Spencer, S. J. (1998). Stereotype threat and the academic underperformance of minorities and women. In J. K. Swim, & C. Stangor (Eds.). *Prejudice* (pp. 83–103). San Diego, CA: Academic Press. https://doi.org/10.1016/B978-012679130-3/50039-9.

Arvate, P. R., Galilea, G. W., & Todescat, I. (2018). The queen bee: A myth? The effect of top-level female leadership on subordinate females. *Leadership Quarterly, 29*(March), 533–548. https://doi.org/10.1016/j.leaqua.2018.03.002.

Azmat, G., & Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*. https://doi.org/10.1016/j.jpubeco.2010.04.001.

Azmat, G., & Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour Economics, 30*, 32–40. https://doi.org/10.1016/j.labeco.2014.06.005.

Babcock, L., Recalde, M., Vesterlund, L., & Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *The American Economic Review, 107*(3), 714–747. https://doi.org/10.1257/AER.20141734.

Bagues, M. F., Sylos-Labini, M., & Zinovyeva, N. (2017). Does the gender composition of scientific committees matter?? *The American Economic Review, 107*(4), 1207–1238. https://doi.org/10.2139/ssrn.2628176.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230–244. https://doi.org/10.1037/0022-3514.71.2.230.

Bass, B. (1985). *Leadership and performance beyond expectations*. New York: The Free Press.

Berge, L. I. O., Juniwaty, K. S., & Sekei, L. H. (2016). Gender composition and group dynamics: Evidence from a laboratory experiment with microfinance clients. *Journal of Economic Behavior and Organization, 131*, 1–20. https://doi.org/10.1016/j.jebo.2016.07.015.

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics, 510*, 1–48. https://doi.org/10.1016/j.jfluchem.2010.02.012.

Born, A., Ranehill, E., & Sandberg, A. (2018). A man's world? The impact of a male dominated environment on female leadership. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3207198.

Burgess, D., & Borgida, E. (1999). Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, Public Policy, and Law, 5*(3), 665–692. https://doi.org/10.1037/1076-8971.5.3.665.

Burns, J. (1978). *Leadership*. New York: Harper & Row.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*(1), 40–48. https://doi.org/10.1177/1745691613513470.

Charness, G., Rigotti, L., & Rustichini, A. (2007). Individual behavior and group membership. *The American Economic Review, 97*(4), 1340–1352.

Charness, G., & Rustichini, A. (2011). Gender differences in cooperation with group membership. *Games and Economic Behavior, 72*(1), 77–85.

Chasteen, A. L., Kang, S. K., & Remedios, J. D. (2011). Aging and stereotype threat: Development, process, and interventions. In M. Inzlicht, & T. Schmader (Eds.). *Stereotype threat: Theory, process, and application.*Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199732449.003.0013.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics, 129*(4), 1625–1660. https://doi.org/10.1093/qje/qju023.Advance.

Cohen, L. L., & Swim, J. K. (1995). The differential impact of gender ratios on women and men: Tokenism, self-confidence, and expectations. *Personality and Social Psychology Bulletin, 21*(9), 876–884. https://doi.org/10.1177/0146167295219001.

Connelly, M. S., Gilbert, J. A., Zaccaro, S. J., Threlfall, K. V., Marks, M. A., & Mumford, M. D. (2000). Exploring the relationship of leadership skills and knowledge to leader performance. *Leadership Quarterly, 11*(1), 65–86. https://doi.org/10.1016/S1048-9843(99)00043-0.

Cota, A. a., & Dion, K. L. (1986). Salience of gender and sex composition of ad hoc groups: An experimental test of distinctiveness theory. *Journal of Personality and Social Psychology, 50*(4), 770–776. https://doi.org/10.1037/0022-3514.50.4.770.

Croizet, J. C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24*(6), 588–594. https://doi.org/10.1177/0146167298246003.

Davies, A. P. C., & Shackelford, T. K. (2008). Two human natures: How men and women evolved different psychologies. *Foundations of evolutionary psychology: Ideas, issues and applications* (pp. 261–280). .

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7*(1), https://doi.org/10.1371/journal.pone.0029081.

Druskat, V. U. (1994). Gender and leadership style: Transformational and transactional leadership in the Roman Catholic church. *The Leadership Quarterly, 5*(2), 99–119. https://doi.org/10.1016/1048-9843(94)90023-X.

Eagly, A. H. (2016). When passionate advocates meet research on diversity, does the honest broker stand a chance? *Journal of Social Issues, 72*(1), 199–222. https://doi.org/10.1111/josi.12163.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573–598. https://doi.org/10.1037/0033-295X.109.3.573.

Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the evaluation of

leaders and meta-analysis. *Psychological Bulletin, 111*(1), 3–22.
Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics, 10*(2), 171–178. https://doi.org/10.1007/s10683-006-9159-4.
Fleishman, E. A., Mumford, M. D., Zaccaro, S. J., Levin, K. Y., Korotkin, A. L., & Hein, M. B. (1991). Taxonomic efforts in the description of leader behavior: A synthesis and functional interpretation. *The Leadership Quarterly, 2*(4), 245–287. https://doi.org/10.1016/1048-9843(91)90016-U.
Flore, P. C., Mulder, J., & Wicherts, J. M. (2019). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology,* 1–35. https://doi.org/10.1080/23743603.2018.1559647.
Freeman, R. B., & Gelber, A. M. (2010). Prize structure and information in tournaments: Experimental evidence. *American Economic Journal: Applied Economics, 2*(1), 149–164. https://doi.org/10.1257/app.2.1.149.
French, J. R. P., & Raven, B. (1968). The bases of social power. In D. Cartwright, & A. F. Zander (Eds.). *Group dynamics: Research and theory* (pp. 259–269). New York: Harper & Row.
Gilardi, F. (2015). The temporary importance of role models for women's political representation. *American Journal of Political Science, 59*(4), 957–970. https://doi.org/10.1111/ajps.12155.
Gneezy, U., & Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review, 94*(2), 377–381. https://doi.org/10.1257/0002828041301821.
Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association.* https://doi.org/10.1007/s40881-015-0004-4.
Haslam, S. A., Ryan, M. K., Kulich, C., Trojanowski, G., & Atkins, C. (2010). Investing with prejudice: The relationship between women's presence on company boards and objective and subjective measures of company performance. *British Journal of Management, 21*(2), 484–497. https://doi.org/10.1111/j.1467-8551.2009.00670.x.
Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues, 57*(4), 657–674. https://doi.org/10.1111/0022-4537.00234.
Heilman, M. E., & Haynes, M. C. (2005). No credit where credit is due: Attributional rationalization of women's success in male-female teams. *Journal of Applied Psychology, 90*(5), 905–916. https://doi.org/10.1037/0021-9010.90.5.905.
Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks?: The implied communality deficit. *Journal of Applied Psychology, 92*(1), 81–92. https://doi.org/10.1088/0305-4470/16/9/026.
Heilman, M. E., & Wallen, A. S. (2010). Wimpy and undeserving of respect: Penalties for men's gender-inconsistent success. *Journal of Experimental Social Psychology, 46*(4), 664–667. https://doi.org/10.1016/j.jesp.2010.01.008.
Heilman, M. E., Wallen, A. S., & Fuchs, D. (2004). Tamkins m. m. *Journal of Applied Psychology*. https://doi.org/10.1037/0021-9010.89.3.416.
Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *The Behavioral and Brain Sciences, 24*(3), 383—403;. http://www.ncbi.nlm.nih.gov/pubmed/11682798 discussion 403-51.
Hoogendoorn, S., Oosterbeek, H., & Praag, M. (2013). The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Science, 59*(7), 1514–1528. http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1120.1.
Hoyt, C. L., & Blascovich, J. (2010). The role of leadership self-efficacy and stereotype activation on cardiovascular, behavioral and self-report responses in the leadership domain. *Leadership Quarterly*. https://doi.org/10.1016/j.leaqua.2009.10.007.
Hoyt, C. L., Johnson, S. K., Murphy, S. E., & Skinnell, K. H. (2010). The impact of blatant stereotype activation and group sex-composition on female leaders. *Leadership Quarterly, 21*(5), 716–732. https://doi.org/10.1016/j.leaqua.2010.07.003.
Hoyt, C. L., & Murphy, S. E. (2016). Managing to clear the air: Stereotype threat, women, and leadership. *Leadership Quarterly*. https://doi.org/10.1016/j.leaqua.2015.11.002.
Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science, 11*(5), 365–371. https://doi.org/10.1111/1467-9280.00272.
Joecks, J., Pull, K., & Vetter, K. (2013). Gender diversity in the boardroom and firm performance: What exactly constitutes a “critical mass?”. *Journal of Business Ethics, 118*(1), 61–72. https://doi.org/10.1007/s10551-012-1553-6.
Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*. https://doi.org/10.1037/0021-9010.87.4.765.
Kanter, R. (1977). Some effects of proportions on group life?: Skewed sex ratios and responses to token women. *American Journal of Sociology, 82*(5), 965–990.
Kanthak, K., & Woon, J. (2015). Women don’t run? Election aversion and candidate entry. *American Journal of Political Science, 59*(3), 595–612. https://doi.org/10.1111/ajps.12158.
Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica, 77*(2), 603–606. https://doi.org/10.3982/ECTA7833.
Karpowitz, C. F., Monson, J. Q., & Preece, J. R. (2017). How to elect more women: Gender and candidate success in a field experiment. *American Journal of Political Science, 61*(4), 927–943. https://doi.org/10.1111/ajps.12300.
Kiefer, A., & Sekaquaptewa, D. (2007). Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility to stereotype threat. *Journal of Experimental Social Psychology, 43*(5), 825–832.
Kirsch, A. (2018). The gender composition of corporate boards: A review and research agenda. *Leadership Quarterly, 29*(2), 346–364. https://doi.org/10.1016/j.leaqua.2017.06.001.
Kray, L. J., Thompson, L., & Galinsky, A. (2001). Battle of the sexes: Gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology, 80*(6), 942–958. https://doi.org/10.1037//0022-3514.80.6.942.
Levy, B. (1996). Improving memory in old age through implicit self-stereotyping. *Journal of Personality and Social Psychology, 71*(6), 1092–1107. https://doi.org/10.1037/0022-3514.71.6.1092.
Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management, 64*, 19–40. https://doi.org/10.1016/j.jom.2018.10.003.
Lord, R. G. (1977). Functional leadership behavior. *Measurement and relation to social power and leadership perceptions. Administrative Science Quarterly, 22*(1), 114–133. https://doi.org/10.2307/2391749.
Miller, T., & Del Carmen Triana, M. (2009). Demographic diversity in the boardroom: Mediators of the board diversity-firm performance relationship. *Journal of Management Studies, 46*(5), 755–786. https://doi.org/10.1111/j.1467-6486.2009.00839.x.
Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). Managing Self-Confidence: Theory and experimental evidence. *NBER working paper series.*https://doi.org/10.1017/CBO9781107415324.004.
Morgeson, F. P., DeRue, D. S., & Karam, E. P. (2010). Leadership in teams: A functional approach to understanding leadership structures and processes. *Journal of Management, 36*. https://doi.org/10.1177/0149206309347376.
Mumford, M. D., & Van Doorn, J. R. (2001). The leadership of pragmatism: Reconsidering Franklin in the age of charisma. *Leadership Quarterly, 12*(3), 279–309. https://doi.org/10.1016/S1048-9843(01)00080-7.
Reitan, T., & Stenberg, S.Å. (2019). From classroom to conscription. Leadership emergence in childhood and early adulthood. *Leadership Quarterly, 30*(3), 298–319. https://doi.org/10.1016/j.leaqua.2018.11.006.
Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology, 74*(3), 629–645.
Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57*(4), 743–762. https://doi.org/10.1111/0022-4537.00239.
Schlag, K. H., Tremewan, J., & van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics, 18*(3), 457–490. https://doi.org/10.1007/s10683-014-9416-x.
Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology*. https://doi.org/10.1016/S0022-1031(02)00508-5.
Shih, M. J., Pittinsky, T. L., & Ho, G. C. (2011). Stereotype boost: Positive outcomes from the activation of positive stereotypes. In M. Inzlicht, & T. Schmader (Eds.). *Stereotype threat: Theory, process, and application.*Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199732449.003.0009.
Smith, N., Smith, V., & Verner, M. (2006). Do women in top management affect firm performance? A panel study of 2,500 Danish firms. *International Journal of Productivity and Performance Management*. https://doi.org/10.1108/17410400610702160.
Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373.
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797.
Stone, J., Sjomeling, M., Lynch, C. I., & Darley, J. M. (1999). Stereotype threat effects on black and white athletic performance. *Journal of Personality and Social Psychology, 77*(6), 1213–1227. https://doi.org/10.1037/0022-3514.77.6.1213.
Terjesen, S., Sealy, R., & Singh, V. (2009). Women directors on corporate boards: A review and research agenda. *Corporate Governance: An International Review, 17*(3), 320–337. https://doi.org/10.1111/j.1467-8683.2009.00742.x.
Tosi, H. L., & Einbender, S. W. (1985). The effects of the type and amount of information in sex discrimination research: A meta-analysis. *Academy of Management Journal, 28*(3), 712–723. https://doi.org/10.2307/256127.
Yang, P., Riepe, J., Moser, K., Pull, K., & Terjesen, S. (2019). Women directors, firm performance, and firm risk. *A causal perspective. The Leadership Quarterly*. https://doi.org/10.1016/j.leaqua.2019.05.004.
Zehnder, C., Herz, H., & Bonardi, J. P. (2017). A productive clash of cultures: Injecting economics into leadership research. *Leadership Quarterly, 28*(1), 65–85. https://doi.org/10.1016/j.leaqua.2016.10.004.
Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics, 13*(1), 75–98. https://doi.org/10.1007/s10683-009-9230-z.